Feature Selection for Anomaly Detection Using Optical Emission Spectroscopy

Feature Selection for Anomaly Detection Using Optical Emission Spectroscopy

Luca Puggini * **Seán McLoone** **

* *National University of Ireland Maynooth, Maynooth, Ireland (e-mail: lpuggini@eeng.nuim.ie)*
** *Queen's University Belfast, Belfast, England (e-mail: s.mcloone@qub.ac.uk)*

Abstract: To maintain the pace of development set by Moore's law, production processes in semiconductor manufacturing are becoming more and more complex. The development of efficient and interpretable anomaly detection systems is fundamental to keeping production costs low. As the dimension of process monitoring data can become extremely high anomaly detection systems are impacted by the curse of dimensionality, hence dimensionality reduction plays an important role. Classical dimensionality reduction approaches, such as Principal Component Analysis, generally involve transformations that seek to maximize the explained variance. In datasets with several clusters of correlated variables the contributions of isolated variables to explained variance may be insignificant, with the result that they may not be included in the reduced data representation. It is then not possible to detect an anomaly if it is only reflected in such isolated variables. In this paper we present a new dimensionality reduction technique that takes account of such isolated variables and demonstrate how it can be used to build an interpretable and robust anomaly detection system for Optical Emission Spectroscopy data.

Keywords: Semiconductors, Fault Detection, Dimensionality Reduction, OC-SVM, OES Spectrum

1. INTRODUCTION

Semiconductor manufacturing is one of the most rapidly evolving industries. To remain competitive producers must continually deliver new devices that are smaller, faster and/or more energy efficient that previous generations, while at the same time keeping production costs low. In this context the ability to detect faults during the production process reduces the number of incorrectly processed wafers and directly translates into improved overall process yield and throughput (He and Wang (2007)).

As a result, fault or anomaly detection is an active area of research within the semiconductor manufacturing environment. Some recent examples are Puggini et al. (2015) and Mahadevan and Shah (2009) where anomaly detection in OES time series is performed with unsupervised random forest and one class support vector machines (OC-SVM) or Ren and Lv (2014), He and Wang (2007) and Verdier and Ferreira (2011) where clustering is used to separate normal and anomaly samples.

Data driven anomaly detection systems can roughly be divided into three subgroups according to the information available about the data during the training phase:

- *Supervised anomaly detection* where samples from normal and abnormal behaving wafers are available to train classifiers such as Linear Discriminant Analysis (LDA), Support Vector Machines (SVM) and k-nearest neighbours to distinguish between normal and anomalous samples (Chandola et al. (2009)).
- *Semi-supervised anomaly detection* where only data for normal samples is available. Systems can then be trained to assign an anomaly score to new samples according to how distant they are from the normal behaving ones. Several algorithms have been developed with this aim, including Multivariate Control Charts (Lowry and Montgomery (1995)), one-class SVMs (Schölkopf et al. (2001)) and Unsupervised Random Forests (Shi and Horvath (2006)).
- *Unsupervised anomaly detection* where no information is available about the data (i.e. the data is unlabeled) but assumptions are made regarding the frequency and distinctiveness of the anomalies within the overall dataset. This structure is then revealed and potential anomalies identified though the application of unsupervised clustering techniques such as DBSCAN (Ester et al. (1996)) and Max Separation clustering (Flynn and McLoone (2011)).

Optical emission spectroscopy (OES) is increasingly being used by semiconductor manufacturers for plasma etch process monitoring due to its ability to track variations in the chemical composition of a plasma over time. The OES data is composed of measurements of the light emitted from the plasma as a function of wavelength and time. Figure 1 shows a sample spectrum from the plasma etch process case study which will be introduced in Section 3. OES has been shown to be an effective wafer processing monitoring signal (Chen et al. (1996), Puggini et al. (2014)) and has been employed for applications such as anomaly detection (Puggini et al. (2015), Yue et al.

(2000)) and etch rate prediction (Puggini and McLoone (2015), Zeng and Spanos (2009)). OES data is generally characterized by high dimension, (Prakash et al. (2012)) which poses a problem for anomaly detection algorithms. Most anomaly detection algorithms are based on a distance measure and it is known that distance measures become meaningless in high dimensional spaces due to the so-called curse of dimensionality (Kriegel et al. (2008)).

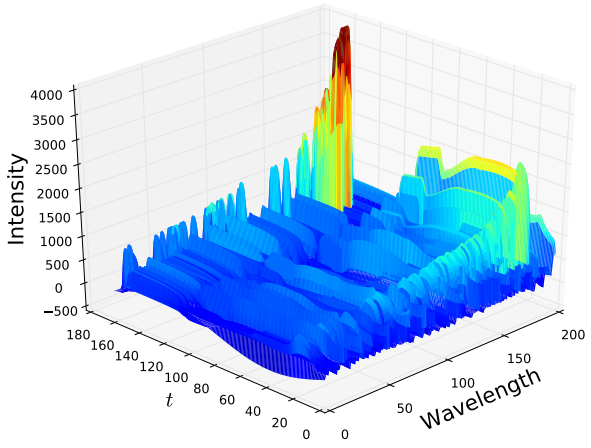

Fig. 1. A typical of OES spectrum from the case study presented in Section 3

In this paper the focus is on developing an appropriate data representation and dimensionality reduction technique for anomaly detection using OES in semiconductor manufacturing. In particular, a Forward Selection Independent Variables (FSIV) algorithm is proposed as an enhancement to Forward Selection Component Analysis (FSCA) (Prakash et al. (2012)) that yields better features for anomaly detection than Principal Component analysis (PCA) (Jolliffe (2002)) or FSCA when the anomaly occurs in an isolated variable in high dimensional correlated datasets. The efficacy of FSIV is demonstrated using both simulated and industrial case studies. In the industrial case study, a semi-supervised anomaly detection system is developed using a one-class SVM as the classification engine.

The remainder of the paper is organised as follows. Section 2 introduces the FSIV algorithm and demonstrates it performance with respect to PCA and FSCA for a simulated example. Similar results ar then presented in Section 3 for an industrial plasma etch case study. The anomaly detection classifier is developed in Section 4 and the results of its application to the industrial case study presented in Section 5. Finally, conclusions are provided in Section 6.

2. DIMENSIONALITY REDUCTION IN ANOMALY DETECTION

Dimensionality reduction techniques such as PCA and FSCA seek to obtain lower dimensional approximations of datasets from which it is possible to reconstruct the majority of the information in the original high dimensional datasets, usually defined in terms the percentage of explained variance. While they are generally very useful for generating compact representations of highly correlated datasets, the reduced representations are not guaranteed to retain sufficient information to detect isolated anomalies. In particular, in datasets with several large clusters of correlated variables, the contributions of isolated uncorrelated variables to explained variance may be insignificant, with the result that such variables may not be included in the reduced data representation. It is then not possible to detect an anomaly if it is only reflected in such isolated variables.

Mitra et al. (2002) and Flynn and McLoone (2011) have developed algorithms that perform unsupervised features selection while at the same time attempting to retain isolated variables in the data. In these algorithms the variables are recursively clustered. In the former for each variable the set of its k-nearest variables is computed according to a similarity function. The variable which is closest to its k^{th} neighbour is retained while its k neighbours are discarded. The process ends when all the k-neighbours of all the variables are closer than a certain threshold to their centroid. In the latter centroids for new clusters are chosen based on how different they are from the data in existing clusters, and individual clusters are formed on the basis of exceeding a similarity threshold. Then when clustering is complete the reduced dataset representation is defined as the centroids of the clusters.

2.1 FSIV Algorithm

Both Mitra et al. (2002) and Flynn and McLoone (2011) select features based on a function $s(x, y)$ that measures the similarity between two variables. In general, instead of discarding variables that are similar to those already selected, it is more interesting to know which variables are not adequately represented by the selected variables. With this in mind Forward Selection Independent Variable (FSIV) analysis is proposed as a tool for efficient unsupervised features selection in anomaly detection.

Here the steps required to select K variables with FSCA (Prakash et al. (2012)) are recalled:

1 Start with the full data $X = (x_1, \ldots, x_p)$ and K the number of variables to select. Initialize $Z_0 = \emptyset$ and $k = 0$.
2 Scale the data to zero mean.
3 Define Z^v_{k+1} as the matrix Z_k with the addition of the variable x_v i.e. $Z^v_{k+1} = (Z_k, x_v)$
4 Define Z_{k+1} as:

$$argmin_v \parallel X - Z^v_{k+1}(Z^T_{k+1^v} Z^v_{k+1})^{-1} {Z^v_{k+1}}^T X \parallel_2 \quad (1)$$

5 Update $k = k + 1$
6 If $k < K$ return to step 3. Otherwise output Z_K, the set of selected variables.

The FSIV algorithm begins by selecting its first k variables $(z_1, \ldots, z_k)$ using the FSCA algorithm. This step is required to ensure the presence of the variables that represent the largest variation in the data. Then, additional variables are added in order to model signficant isolated variations that are not captured by the first k variables. The process ends when K variables are selected or when the error $\epsilon_{\hat{j}}$ defined according to equations 4 and 5 is smaller than a given threshold. The FSIV algorithm is thus defined as follows:

1 Start with the full data $X = (x_1, \ldots, x_p) \in \mathbb{R}^{n \times p}$ and set k and a stop criterion.
2 Scale the data such that each variable has zero mean.
3 Select k variables $z_1, \ldots, z_k$ using the FSCA algorithm.
4 Define the matrix $Z = (z_1, \ldots, z_k)$.
5 Compute the linear approximation of X

$$\hat{X} = Z(Z^T Z)^{-1} Z^T X \in \mathbb{R}^{n \times k} \tag{2}$$

where

$$\hat{X} = (\hat{x}_1, \ldots, \hat{x}_p) \tag{3}$$

6 For each variable x_i in X compute its approximation error

$$\epsilon_i = \| x_i - \hat{x}_i \|_2^2 \tag{4}$$

where $\hat{x}_i$ is the ith column of $\hat{X}$.
7 Select $x_{\hat{j}}$ the variable with the highest approximation error where:

$$\hat{j} = \underset{i}{\operatorname{argmax}}\, \epsilon_i \tag{5}$$

8 Add $x_{\hat{j}}$ to the Z matrix.
9 Stop if the termination criterion is reached, otherwise set $k = k + 1$ and repeat from step 5

The distinguishing feature of FSIV is that a variable is added to the model if it cannot be adequately reconstructed by a linear combination of those already selected. This makes the algorithm more efficient than methods based on similarity between variables. This follows, for example, from the fact that lowly correlated variables may be linearly dependent (Rodgers et al. (1984)). The following example illustrates the difference in performance between PCA, FSCA and FSIV.

2.2 Simulated Example

Consider the simulated data $X = (x_1, \ldots, x_7) \in \mathbb{R}^{n \times 7}$ defined by three groups of variables $X_1 = \{x_1, x_2, x_3\}$, $X_2 = \{x_4, x_5, x_6\}$ and $X_3 = \{x_7\}$. Each variable has correlation 0.9 with the others in the same group and between the variables in X_1 and X_2 there is a correlation of 0.4. The variable in X_3 is instead isolated and has only correlation 0.1 with all other variables. Specifically, $X = (x_1, \ldots, x_7) \sim N(0, \Sigma)$ where $\Sigma = \{\Sigma_{i,j}\}$ is defined as:

$$\Sigma_{i,j} = \begin{cases} 1 & if\ i = j \\ 0.9 & if\ i,j \in \{1,2,3\}\ or\ i,j \in \{4,5,6\} \\ 0.4 & if\ i \in \{1,2,3\}\ and\ j \in \{4,5,6\} \\ 0.4 & if\ j \in \{1,2,3\}\ and\ i \in \{4,5,6\} \\ 0.1 & if\ i = 7\ or\ j = 7 \end{cases}$$

An anomaly is then introduced by replacing one of the samples in x_7 with the value 10. Dimensionality reduction is performed with PCA, FSCA and FIV. In each case only two variables are selected. In FSIV parameter k is chosen as $k = 1$. The two dimensional representations of the data obtained with the various methods is reported in Figure 2. From the figure it can be observed that only FSIV is able to isolate the anomaly. In particular, FSCA tends to select one variable from X_1 and one from X_2 while FSIV selects a variable from X_1 and x_7. The PCA components instead are obtained as a weighted linear combination of all the variables. However, the weighting associated with x_7 is insufficient to materially affect the behaviour of the components, with the result that the anomaly is not distinguishable from the normal samples.

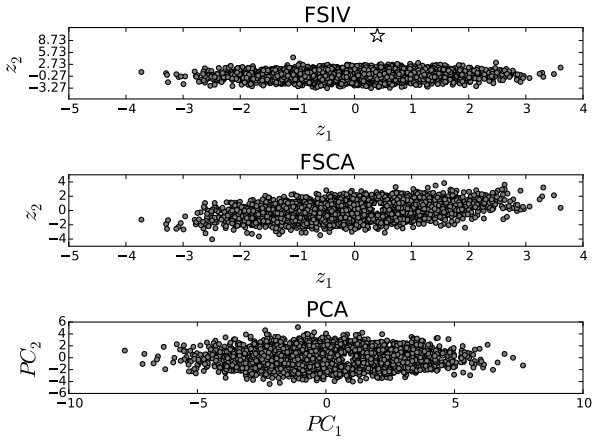

Fig. 2. The projection of the data for the simulated example on the first two variables selected by FSCA and FSIV, and the first two principal components obtained with PCA

3. INDUSTRIAL CASE STUDY

To demonstrate the effectiveness of the proposed dimensionality reduction method, the technique is applied to a sample dataset from an industrial plasma etch chamber. The dataset consists of Plasma Etch Optical Emission Spectroscopy (OES) samples.

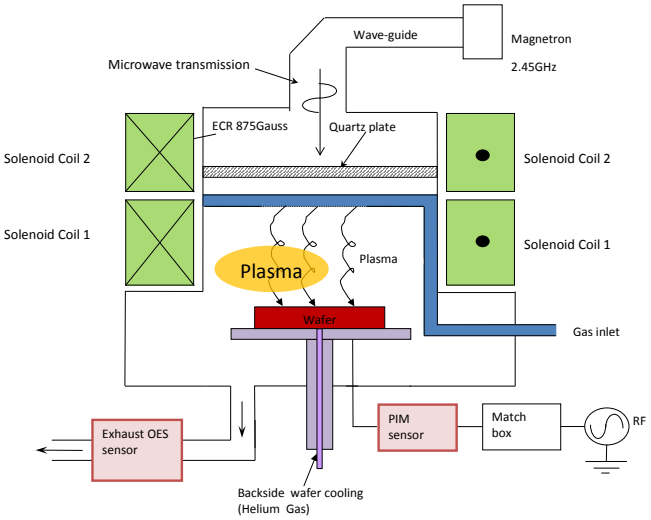

Fig. 3. A plasma etching chamber.

OES recordings were taken from the etching chamber exhaust, as depicted in Fig. 3. By using only OES data for process monitoring, the proposed dimensionality reduction and related fault-detection methodology is able to operate in real-time, thereby reducing the risk of costly faults propagating during production.

3.1 OES data

Noting that the OES data is naturally parameterized in terms of the wafer number, the processing time instant and

the measured wavelengths (Yue et al. (2000)), the intensity of the i^{th} wavelength of the k-th wafer at time t is denoted as $x_i^{w_k}(t)$.

OES spectra for $K = 500$ wafers are available, each one consisting of $\tau = 165$ samples of $p = 200$ wavelengths. The OES spectrum for a single wafer w_k can be mathematically represented as a matrix $X_k \in \mathbb{R}^{\tau \times p}$.

$$X_k = \left\{ x_i^{w_k}(t_{(k-1)\tau+j}) \right\}_{j=1,\ldots,\tau,\ i=1,\ldots,p} \in \mathbb{R}^{\tau \times p} \qquad (6)$$

and the full data is represented by a set S containing the measurements for each wafer:

$$S = \left\{ X_j \in \mathbb{R}^{t_\tau \times p} :\ j = 1, \ldots, K \right\} . \qquad (7)$$

For practical purposes it is better to store the data in a two dimensional matrix. Two possible aggregations are considered and are denoted as $\Lambda \in \mathbb{R}^{\tau K \times p}$ and $W \in \mathbb{R}^{K \times p\tau}$. These will be discussed in sections 3.2 and 3.4, respectively.

Artificial Fault In order to better show the difference between FSCA and FSIV an artificial wavelength $x_l^{w_k}(t)$ is added to the OES spectrum. $x_l^{w_k}(t)$ is defined for each wafer as

$$x_l^{w_k}(t) = 285(sin(t) + \epsilon) \quad t \in [-\pi, \pi] \qquad (8)$$

where $\epsilon \sim N(0, 0.05)$ and amplitude 285 is selected to give a signal power that is similar to the other wavelengths. A fault is then introduced in the final wafer in the dataset by clamping the l^{th} wavelength to lie between -100 and 100, that is:

$$x_l^{w_K}(t) = \begin{cases} x_l^{w_K}(t) & if\ |x_l^{w_K}(t)| < 100 \\ 100 + \epsilon & if\ x_l^{w_K}(t) > 100 \\ -100 + \epsilon & if\ x_l^{w_K}(t) < -100 \end{cases} \qquad (9)$$

where $\epsilon \sim N(0, 10)$. In figure 4 the artificial wavelength $x_l^{w_k}(t)$ and a normal wavelength are illustrated for a group of five wafers. It follows that the faulty wafer can be classified as anomaly only if the wavelength l is among the selected ones.

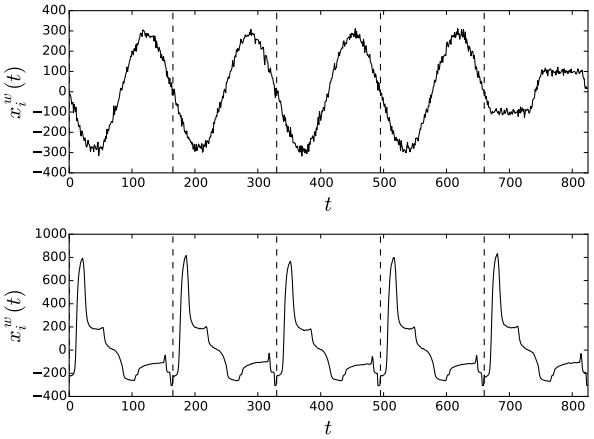

Fig. 4. The artificial wavelength and a normal wavelength over 800 time points and 5 wafers.

3.2 The Λ matrix

The data can be aggregated in a $\Lambda \in \mathbb{R}^{\tau K \times p}$ matrix. In Λ each row corresponds to a time scan and each column to a wavelength. Λ can be obtained by vertically stacking the matrices in S.

$$\Lambda = \begin{pmatrix} X_1 \\ X_2 \\ \cdots \\ X_K \end{pmatrix} \in \mathbb{R}^{K\tau \times p} \qquad (10)$$

The full OES data is then represented as a set of matrices where each one contains the spectrum for a given wafer. This format of the data is used to select a subset of relevant wavelengths. The wavelengths are chosen with the FSIV and FSCA algorithms.

3.3 Data approximation with FSCA and FSIV

FSCA and FSIV select variables using different criteria. Figure 5 shows the maximal error $\epsilon_{\hat{j}}$ defined according to equations 4 and 5, as a function of the number components selected by FSCA and FSIV, while figure 6 shows the corresponding explained variance (EV). In total 14 components are selected by both the FSCA and FSIV algorithms. The first 7 components of the FSIV algorithm are selected with FSCA, hence it follows that the performances of the both methods in terms of both EV and $\epsilon_{\hat{j}}$ are identical for these components. In contrast for the remaining 7 components we can observe that as expected the variables selected with FSIV lead to a lower error $\epsilon_{\hat{j}}$, while those selected by FSCA yield a larger percentage of EV. Notably, the l^{th} wavelength is not selected by FSCA but is selected by FSIV as the 8^{th} component. It can be observed in Figure 5 that the 8^{th} component is where the performance of FSCA and FSIV begin to deviate in terms of the error $\epsilon_{\hat{j}}$.

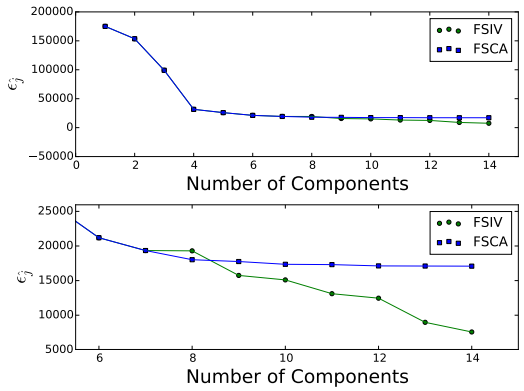

Fig. 5. The error $\epsilon_{\hat{j}}$ as a function of the number of components selected with FSCA or FSIV.

3.4 W Matrix

Alternatively the data can be aggregated in order to have each wafer as an observation. For each wafer all the time scans of all the wavelengths are stored in a row. This is equivalent to transforming all the matrices in S into

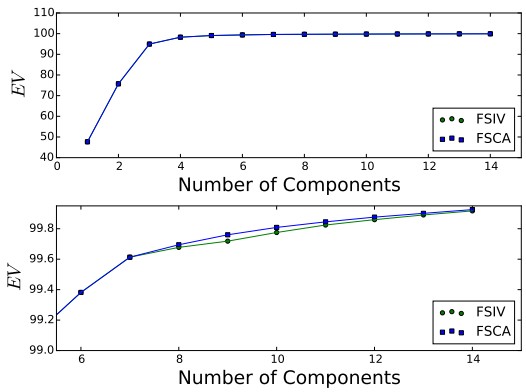

Fig. 6. The error percentage of explained variance as a function of the number of components selected with FSCA or FSIV. The vertical line represent the value of k for the FSIV algorithm.

vectors and stacking them horizontally. Matrix X_k, the k_{th} element of S is reshaped as

$$\tilde{X}_k = (x_1^{w_k}(t_1), \ldots, x_p^{w_k}(t_\tau)) \in \mathbb{R}^{1\times\tau p} \tag{11}$$

and the full dataset is represented by combining all the reshaped matrices in S as:

$$W = \begin{pmatrix} \tilde{X}_1 \\ \cdots \\ \tilde{X}_K \end{pmatrix} \in \mathbb{R}^{K\times\tau p} \tag{12}$$

This data format is particularly useful for comparing wafers and performing anomaly detection as each row corresponds to all the data observed for a given wafer. However, the dimension of the matrix is very large as it has $p\tau = 33000$ columns. The dimension of the data can be drastically reduced by selecting only a subset of the wavelengths from the Λ matrix. If 14 wavelengths are selected from Λ the number of columns in the W matrix is reduced to $14\tau = 2310$ columns. The W matrix is still high dimensional but now it is small enough for high dimensional anomaly detection algorithms such as Unsupervised Random Forest and OC-SVM to be efficiently applied.

4. OC-SVM ANOMALY DETECTION

Once a lower dimensional approximation of the W matrix is obtained it is possible to use it to develop an anomaly detection system. In plasma etching wafers are normally processed is batches of 25, called a lot, with the chamber undergoing a cleaning cycle between each lot. As a consequence of the cleaning step there is a seasoning effect during the processing of the first few wafers in each lot as chemicals absorb into the chamber walls, with the result that the processing of wafers $1, 2, 3$ and 4 differ slightly from the remaining wafers 5 to 25. For the purposes of evaluating the performance of the different dimensionality reduction techniques as a per-processing step for anomaly detection, we will consider wafers $1, 2, 3$ and 4 in each lot as abnormal wafers. In addition, wafer 500 is also defined as abnormal due to the artificial anomaly introduced in the l^{th} wavelength.

In order to train and test the anomaly detector the wafers in W are split into a training set of 300 wafers containing measurements of only normal behaving wafers and a test set of 200 wafers containing normal and abnormally behaving wafers. The OC-SVM algorithm is used to assign an anomaly score to each wafer.

Given a training dataset $X \in \mathbb{R}^{n\times p}$, $X \subset \mathcal{X}$ where $\mathcal{X}$ is a compact subset of $\mathbb{R}^p$ and Φ a map into the dot product space:

$$\Phi(X) : \mathcal{X} \to F \;\; and \;\; \Phi(x)\cdot\Phi(y) = k(x,y)\; \forall x, y \in \mathcal{X} \tag{13}$$

the OC-SVM optimization problem is defined as:

$$\min_{w\in F, \xi\in\mathbb{R}^n, \rho\in\mathbb{R}} \frac{1}{2} \parallel w \parallel + \frac{1}{vn}\sum_{i=1}^{n}\xi_i - \rho \tag{14}$$

subject to the constraint

$$(w\cdot\Phi(x_i)) \geq \rho - \xi_i \quad i = 1, \ldots, n \quad \xi_i \geq 0 \tag{15}$$

Since the ξ are penalized it is expected that the decision function

$$f(x) = sign((w\cdot\Phi(x)) - \rho) \tag{16}$$

will be positive for most samples x_i, when w and ρ are optimized. At the same time $\parallel w \parallel$ is small forcing $f(x) > 0$ only on a small region.

The OC-SVM is trained using only normal behaving wafers. Then an anomaly score, defined as

$$s(x) = -(w\cdot\Phi(x)) + \rho \tag{17}$$

is assigned to each sample x in the test set. In other words the anomaly score is the distance between x and the estimated support of the normal behaving data.

5. RESULTS

The anomaly score assigned by OC-SVM to each wafer in the test dataset when using each of the dimensionality reduction techniques is given in Figure 7. For completeness, the results obtained without dimensionality reduction are also reported. The results show that in general a larger anomaly score is assigned to the abnormal wafers allowing them to be identified. The one exception is the artificially created abnormal wafer, denoted by the star, which is only correctly identified as an anomaly when FSIV is used. Even when all the wavelengths are used the artificial anomaly has a low anomaly score. This may be due to over fitting caused by the excessive number of variables. The performance of each method is also summarized in terms of the Area Under the Curve (AUC) classifier performance metric in Table 5.1 and again underscores the superiority of FSIV for this application.

	A.W.	FSCA	FSIV	PCA
AUC	0.9564	0.9507	0.9650	0.9527

Table 5.1: The AUC score obtained using OC-SVM when all the wavelengths are used (A.W.), when a subset of 14 wavelengths is selected with FSCA and FSIV, and when 14 PCA components are employed as inputs.

6. CONCLUSION

This paper considers the problem of feature selection for anomaly detection with application to OES based semi-

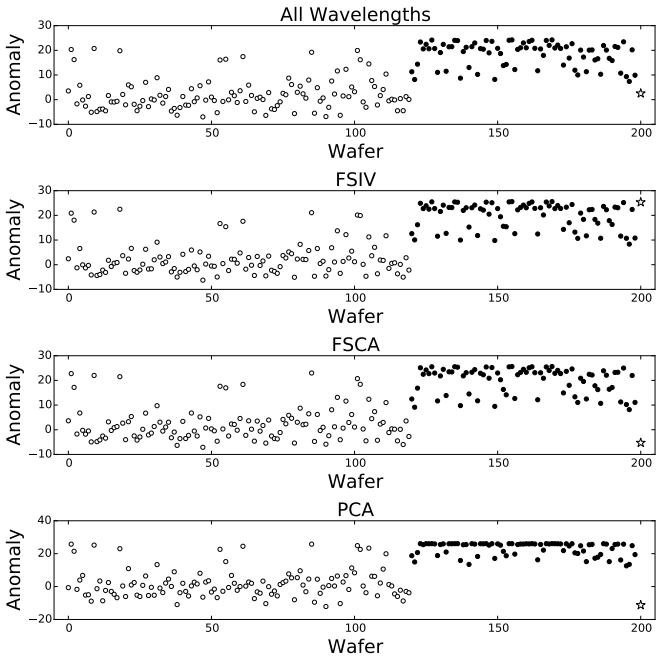

Fig. 7. Anomaly score assigned to each wafer in the test set: Black circles denote the anomaly wafers ($1-4$ in each lot), white circles denote the normal wafers, and the artificial anomaly wafer is represented by a star.

conductor manufacturing process monitoring. FSIV is proposed as a new feature selection method that takes account of isolated variables in highly correlated high dimension datasets. FSIV combined with OC-SVM is evaluated using an industrial case study and shown to outperform FSCA and PCA for anomly detection. While a good separation is achieved between normal and abnormal wafers some false positives are still present. Further research is required to understand the nature of these false positives.

ACKNOWLEDGEMENTS

The authors would like to thank Intel Ireland for providing the industrial case study for this research and Maynooth University for the financial support provided.

REFERENCES

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3), 15.

Chen, R., Huang, H., Spanos, C., and Gatto, M. (1996). Plasma etch modeling using optical emission spectroscopy. *Journal of Vacuum Science & Technology A*, 14(3), 1901–1906.

Ester, M., Kriegel, H.P., Sander, J., and Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, 226–231.

Flynn, B. and McLoone, S. (2011). Max separation clustering for feature extraction from optical emission spectroscopy data. *Semiconductor Manufacturing, IEEE Transactions on*, 24(4), 480–488.

He, Q.P. and Wang, J. (2007). Fault Detection Using the k-Nearest Neighbor Rule for Semiconductor Manufacturing Processes. *IEEE Transactions on Semiconductor Manufacturing*, 20(4), 345–354. doi:10.1109/TSM.2007.907607.

Jolliffe, I. (2002). *Principal component analysis*. Wiley Online Library.

Kriegel, H.P., Zimek, A., et al. (2008). Angle-based outlier detection in high-dimensional data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 444–452. ACM.

Lowry, C.A. and Montgomery, D.C. (1995). A review of multivariate control charts. *IIE transactions*, 27(6), 800–810.

Mahadevan, S. and Shah, S.L. (2009). Fault detection and diagnosis in process data using one-class support vector machines. *Journal of Process Control*, 19(10), 1627–1639. doi:10.1016/j.jprocont.2009.07.011.

Mitra, P., Murthy, C., and Pal, S.K. (2002). Unsupervised feature selection using feature similarity. *IEEE transactions on pattern analysis and machine intelligence*, 24(3), 301–312.

Prakash, P., Johnston, A., Honari, B., and McLoone, S. (2012). Optimal wafer site selection using forward selection component analysis. In *Advanced Semiconductor Manufacturing Conference (ASMC), 2012 23rd Annual SEMI*, 91–96. IEEE.

Puggini, L., Doyle, J., and McLoone, S. (2014). Towards multi-sensor spectral alignment through post measurement calibration correction.

Puggini, L., Doyle, J., and McLoone, S. (2015). Fault detection using random forest similarity distance. *IFAC-PapersOnLine*, 48(21), 583–588.

Puggini, L. and McLoone, S. (2015). Extreme learning machines for virtual metrology and etch rate prediction. In *Signals and Systems Conference (ISSC), 2015 26th Irish*, 1–6. IEEE.

Ren, L. and Lv, W. (2014). Fault Detection via Sparse Representation for Semiconductor Manufacturing Processes. *IEEE Transactions on Semiconductor Manufacturing*, 27(2), 252–259. doi:10.1109/TSM.2014.2302011.

Rodgers, J.L., Nicewander, W.A., and Toothaker, L. (1984). Linearly independent, orthogonal, and uncorrelated variables. *The American Statistician*, 38(2), 133–134.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., and Williamson, R.C. (2001). Estimating the support of a high-dimensional distribution. *Neural computation*, 13(7), 1443–1471.

Shi, T. and Horvath, S. (2006). Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics*, 15(1).

Verdier, G. and Ferreira, A. (2011). Adaptive Mahalanobis Distance and k-Nearest Neighbor Rule for Fault Detection in Semiconductor Manufacturing. *IEEE Transactions on Semiconductor Manufacturing*, 24(1), 59–68. doi:10.1109/TSM.2010.2065531.

Yue, H.H., Qin, S.J., Markle, R.J., Nauert, C., and Gatto, M. (2000). Fault detection of plasma etchers using optical emission spectra. *Semiconductor Manufacturing, IEEE Transactions on*, 13(3), 374–385.

Zeng, D. and Spanos, C.J. (2009). Virtual metrology modeling for plasma etch operations. *Semiconductor Manufacturing, IEEE Transactions on*, 22(4), 419–431.